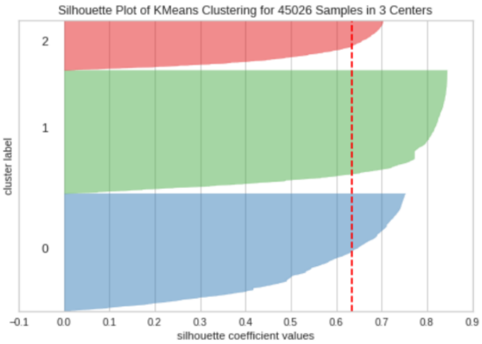
My Colab link: https://colab.research.google.com/drive/1zrb84ZQdT-hNSMI-BKXVL-XBPXD9tzXz

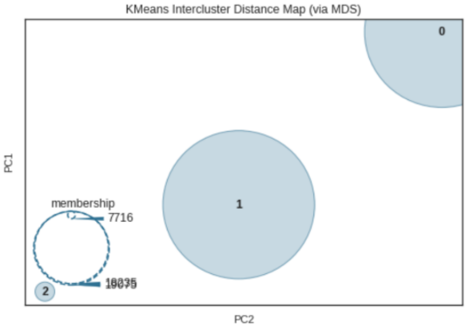
Overview

In this assignment, I make some modifications to the "world food factors feature" notebook. I continue testing the efficiency of the current clusters and try to name them based on their features.

Technique

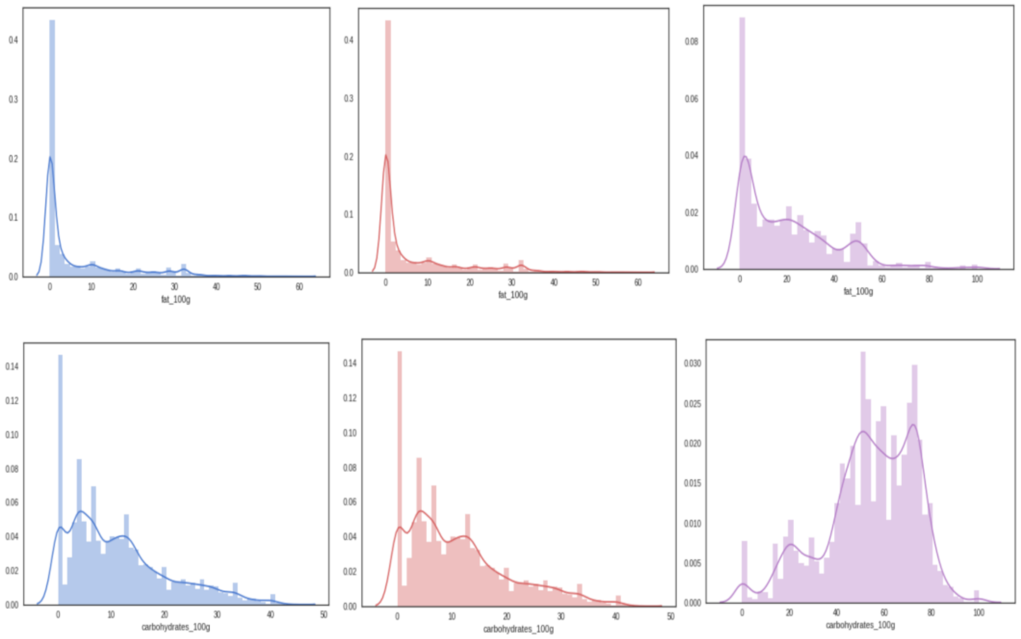
In cluster diagnostics part, the original notebook uses Yellowbrick Silhouette Visualizer to exam the clustering with Minibatch KMeans, I change the model to KMeans, which consists of the former model. Then, I get the following graph.

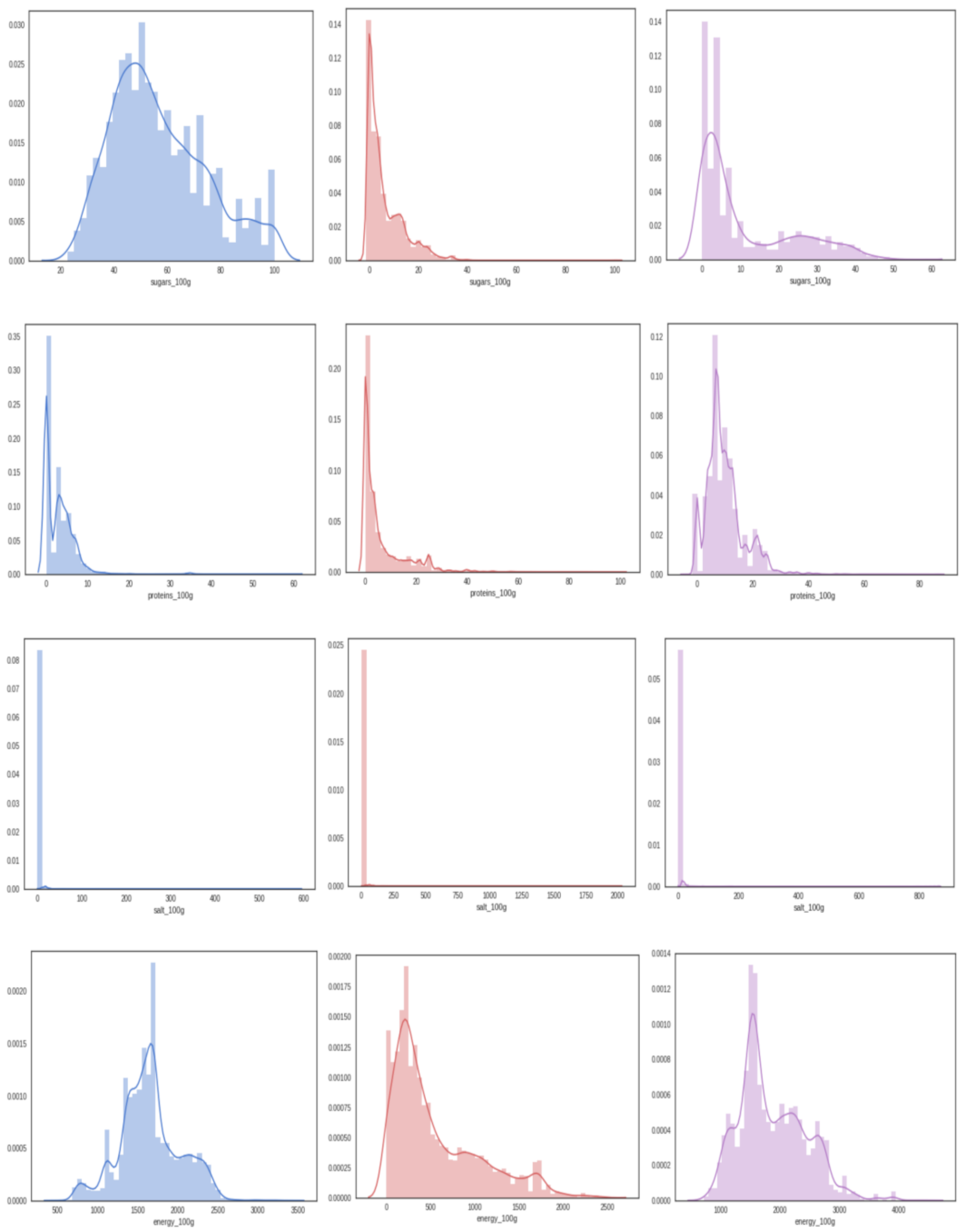




From this graph, we can see all three groups' silhouette coefficient values are greater than 0.7, which means the current clusters are efficient to reflect the features of our data. Then, I draw an inter-cluster distance map based on our clustering, which can display an embedding of the cluster centers in two dimensions with the distance to other centers preserved. Observing from the distance map, we can find foods are mainly concentrated in cluster 0 and cluster 1. All three clusters do not overlap in the original feature space.

To name these three clusters, I describe the distributions of all features in each cluster and try to name the cluster based on its domain features.





Conclusion

By observing and comparing the distribution of features in different clusters, I find that the distribution of sugar content in cluster 0 was significantly different from the other two clusters, so I named it ***high sugar food***. Each nutrient feature of cluster 1 is low, so it can be named as ***low-energy food***. In cluster 2, carbohydrates show obvious characteristics, so it can be named ***high-carb food***.